Building a Parallel Corpus on the World's Oldest Banking Magazine

Volk, Martin ; Amrhein, Chantal ; Aepli, Noëmi ; Müller, Mathias ; Ströbel, Phillip

Abstract: We report on our processing steps to build a diachronic parallel corpus based on the world's oldest banking magazine. The magazine has been published since 1895 in German, with translations in French and partly in English and Italian. Our data sources are printed issues (until 1997), PDF issues (since 1998) and HTML files (since 2001). The corpus building poses special challenges in article boundary recognition and cross-language article and sentence alignment. Our corpus fills a gap in parallel corpora with respect to genre (magazine articles), domain (banking and economy articles), and its time span (120 years).

Building a Parallel Corpus on the World's Oldest Banking Magazine

Martin Volk **Chantal Amrhein** **Noëmi Aepli** **Mathias Müller** **Phillip Ströbel**

University of Zurich
Institute of Computational Linguistics
volk@cl.uzh.ch

Abstract

We report on our processing steps to build a diachronic parallel corpus based on the world's oldest banking magazine. The magazine has been published since 1895 in German, with translations in French and partly in English and Italian. Our data sources are printed issues (until 1997), PDF issues (since 1998) and HTML files (since 2001). The corpus building poses special challenges in article boundary recognition and cross-language article and sentence alignment. Our corpus fills a gap in parallel corpora with respect to genre (magazine articles), domain (banking and economy articles), and its time span (120 years).

1 Introduction

Translated documents in multiple languages (parallel corpora) are highly regarded as valuable resources for natural language processing and linguistic research. But the genres of available parallel corpora are limited to the proceedings of multilingual parliaments (e.g. Europarl), law collections of international bodies (e.g. Acquis Communautaire), subtitles and transcripts (e.g. OpenSubtitles or TED talks) and software manuals. Our goal is to complement these collections with corpora in new genres, domains and a diachronic dimension.

Towards this goal we are building a corpus on the basis of the world's oldest banking magazine, the Credit Suisse Bulletin, which has been published since 1895. The texts in this publication series revolve around all banking issues such as investments, savings, stock prices, but also cover a wide range of topics such as sports, traveling and culture. Over the years, the Bulletin has developed into a full-fledged magazine, currently being published 5 times per year with around 80 pages each in four languages: English, French, German and Italian. Thus we are compiling a parallel corpus in a text genre (magazine articles) that is otherwise seldomly translated. Eventually our corpus will cover a period of more than 120 years.

In addition to the Bulletin, Credit Suisse publishes web news some of which are summaries of the articles in the Bulletin. We will include these news in the corpus since they are valuable resources for comparison.

This paper presents our design decisions for the Credit Suisse Bulletin corpus. First, we introduce the source documents for our collection, then we describe our processing steps and output format, and finally we discuss the challenges in this undertaking. In particular, we propose an algorithm to disambiguate lemmas based on the parallel documents. Overall, this paper focuses on sharing experiences in building a parallel diachronic corpus.

2 Corpus Sources

The Credit Suisse Bulletin has been published since 1895. The copies from the start until 1997 are available only as printed and bound journals in the Credit Suisse archives and some libraries. Scanning bound books is a time-consuming endeavour which requires either manual page turning or expensive scan robots. Therefore, we initially decided that we will use only those copies that we are allowed to cut at the spine for scanning with automatic paper feed. The Credit Suisse archive donated their duplicates from the years 1981 to 1996, in total 262 magazine issues in the four languages English, French, German and Italian. In addition, the library of the Swiss National Bank has offered their issues for cutting, scanning and rebinding, and the Swiss National Library in Berne has agreed to digitize the remaining copies for us.

Second, the Bulletin issues from 1998 to date are available as PDF documents from the Credit

Suisse website[1]. We have downloaded a total of 396 Bulletins as PDF files from which we extracted text content and document structure.

In neither the printed issues nor the PDF issues all articles are translated into all four languages. French and German are mostly given, while English and Italian are sometimes missing. This leads to interesting challenges in document alignment.

The third source of documents for our corpus are the news that are published as HTML documents on the bank's web page. Many of them represent modified versions of articles that were published in the magazine. We crawled all these news (from 2001 to date) which amount to roughly 1500 articles (1.7 million tokens) per language. With a cleaning script, we extracted the text and annotated it with XML markup for title, author, date, and category (such as economy, entrepreneurs, investing, Switzerland).

3 Steps in Corpus Building

The initial steps in corpus building differ depending on the corpus sources.

3.1 Converting Scanned Documents

We converted our scanned Bulletin issues into text with the Abbyy Recognition Server[2]. This OCR program outputs a detailed layout XML with character level information of the coordinates, the system's recognition confidence and the font size as well as word-level information on whether the word is in the OCR system's internal lexicon. The page position coordinates allow us to ignore text in header or footer lines which we do not want in the corpus. They also allow us to detect the page numbers. The font size provides basic information for the detection of the article boundaries. Subsequently, we convert and tokenize the Abbyy XML output into an intermediate XML format.

OCR leads to a small word error rate. For the period which we have digitized so far (1981 to 1996), the recognition accuracy is very high (in text blocks we found less than one incorrect letter in 1000 letters; there are more errors in the occasional words on images). We expect the error rate to be somewhat higher for older Bulletin issues, not least because of more words outside the OCR system's lexicon. If necessary, we will employ the correction methods that we developed in previous projects (Volk et al., 2011). In addition, we consider new methods for word error corrections based on automatic word alignment across the different language versions.

[1] http://publications.credit-suisse.com/index.cfm/publikationen-shop/bulletin/

[2] We gratefully acknowledge support by Abbyy GmbH.

3.2 Converting PDF Documents

There are a number of tools for the extraction of text from PDF documents. Some are freely available, others are part of commercial tools such as Adobe Acrobat. We found that many of those tools deliver ill-formed text, which in our case meant words that were glued together (e.g. *JapansWirtschaft, ein radikalanderes Schulsystem*). After a thorough evaluation we purchased PDFLib TET since it gave the best results and outputs layout XML in a format similar to Abbyy's Recognition Server. Again we get character coordinates and font sizes.

We had hoped that font sizes for article headers in the Bulletin are consistent for certain publication periods. Unfortunately, font sizes for titles are neither consistent nor unique. They differ from year to year and from issue to issue. Sometimes they even differ within the different language versions of the same issue. For example, if an English title is short and set in font size 48, it happens that the corresponding German title is longer and therefore decreased to font size 44 in order to fit the space on the page. Moreover, a large font does not always mean an article headline. We found cases where the same large font is part of an illustration or an advertisement. With this approach the precision of article boundary detection varies from 75% to 95% while recall is as low as 65% and 70% respectively.

3.3 Article Boundary Recognition

Since our results for automatic title detection using only font sizes were not satisfactory, we performed article boundary recognition based on the table-of-content. While testing our prototype system, we realized that the layouts of the individual issues differ more than expected. Sometimes advertisement pages were not included in the page numbering. Therefore, it was not possible to work with incremental page counts, but rather the page numbers had to be extracted directly from the page. Unfortunately, in some years the page numbers where located at the bottom of the pages and in others at the top. Additionally, for a small number of issues the table of contents was stretched across two non-consecutive pages, while in some special issues

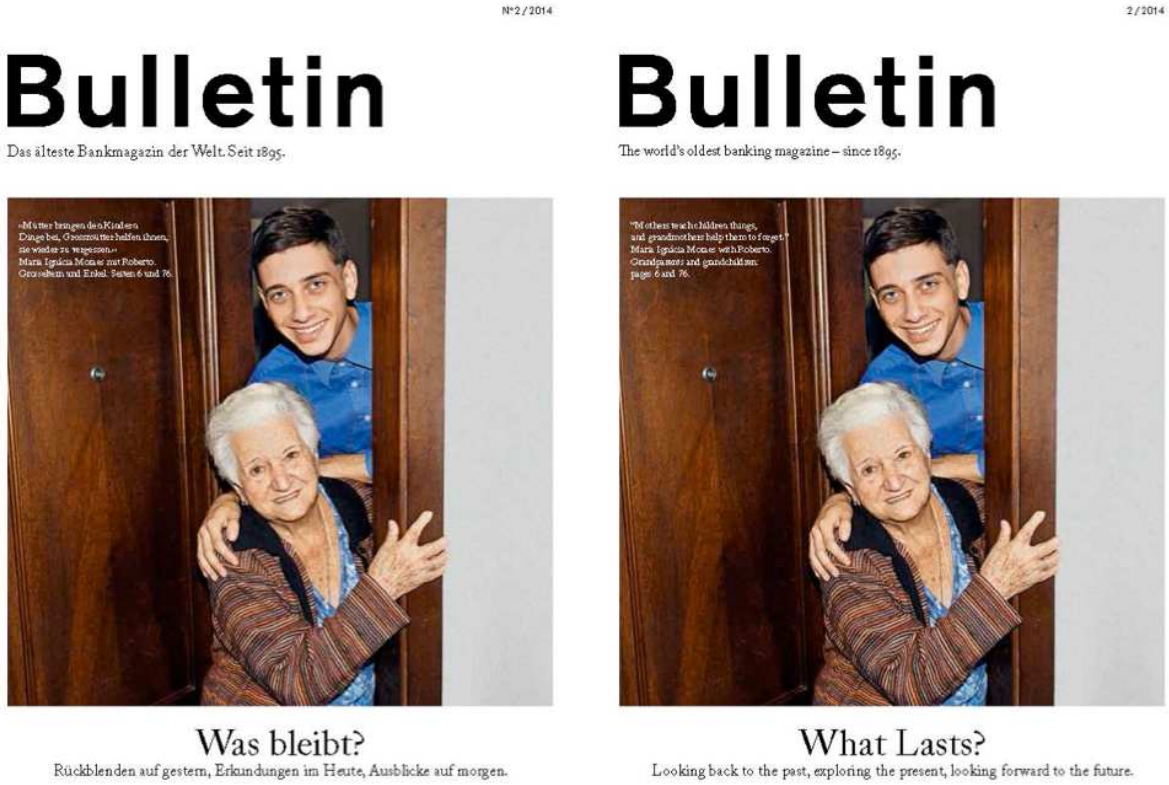

Figure 1: Bulletin 2014, number 2, title page in German and English

the table of contents was missing entirely. Thus, our system required high adaptability to individual layouts.

Another problem arose because of the differences in wording between the article listing in the table of contents and the actual titles in the magazines. For example, we find "Family and Career. Six Women Show How It's Done" in the table of contents of the English version, but the title of the article is "The Art of Compromise". For the majority of articles, it was not possible to match the strings from the table of contents and the article headers. We cannot use this information to identify article boundaries. Moreover, it happens that article titles span over two pages, and that they are integrated in images which led to errors in the text extracted by PDFLib TET. For this reason, we provide the option to confirm or reject the title candidates suggested by our system.

This is how it works: First, we automatically analyze the table-of-content page(s) in order to extract the page numbers where a new article starts. We navigate to the corresponding pages either with an offset, if pagination is continuous, or with page numbers extracted at the top or bottom of the pages. We then search for title candidates. A font size threshold is used to identify large segments on the page. The resulting candidates can be manually confirmed or rejected. If the system does not find an article header, e.g. if it is part of an image, there is also an option to mark a starting article at the beginning of the page. It is possible to extract the article headers fully automatically, however the results will not be as good.

Precision		Recall	
Range	Issues	Range	Issues
0.5 - 0.6	-	0.5 - 0.6	3
0.6 - 0.7	-	0.6 - 0.7	3
0.7 - 0.8	9	0.7 - 0.8	2
0.8 - 0.9	39	0.8 - 0.9	43
0.9 - 1.0	115	0.9 - 1.0	89
1.0	207	1.0	228
avg: 0.961	sum: 370	avg: 0.964	sum: 370

Table 1: Article Boundary Detection Quality: Rows 1 to 6 show the distribution of precision and recall. The last row presents the average results for precision and recall as well as the total number of issues evaluated

Since we achieved satisfactory results when testing our approach on a small number of magazine issues, we performed a large-scale evaluation while we used the system to detect and mark the article boundaries of all PDF issues. If one of the proposed headers was confirmed to be an actual article title, we counted it as a true positive; if none of the proposed candidates matched the actual title, they were collectively counted as one false positive. The number of false negatives was calculated by subtracting the number of found articles from the total number of articles in the table of contents.

Table 1 shows that the F-measure for the majority of articles is between 0.9 and 1.0. These are extremely good results compared to our initial experiments, when we only used font sizes to identify article boundaries. Of course, one has to keep in mind that the evaluation includes a semi-automatic check of the article titles. Without this step the quality would clearly be lower. However, manually checking the proposed candidates does not take much time and is therefore recommended.

Our approach is related to techniques described in (Dejean, 2015) for identifying specific data fields in architectural plans. The tasks are similar in that we also work with unstructured text and use layout characteristics such as font size and text position on the page in order to identify potential article headings. However, unlike content tagging presented in (Dejean, 2015), we do not combine our layout analysis with textual information, for example by making use of the descriptions in the content page or measuring the length of a title candidate. For our task, we achieve very good results by only using the layout. Additionally, we do not generate a data model for every magazine individually. Instead, we use one font size threshold for all magazines, which can of course be adapted for individual issues if needed.

3.4 Corpus Size of the PDF Magazines

All the different corpus sources are stored in an intermediate XML format. To build our parallel corpus, we tokenized and tagged all files. The current size of the Bulletin corpus based on PDF documents is displayed in table 2. For German, French and Italian we have more than 3 million tokens per language. For English we have close to 2 million tokens because English translations have only recently become available for the PDF series.

As for lemmas, table 3 shows the number of unique lemma types as well as the absolute number of tokens whose lemma is unknown for all languages. This should be read as follows: The German corpus which was built on the Bulletin PDF magazines has 3.48 million tokens which account for 201,611 different types (leaving upper case untouched). Out of these 3.48 million tokens we were unable to compute a lemma for 161,644 tokens. Those were unknown to our dictionary and to our morphology analyzer (e.g. loan words like *E-commerce*, names like *Calderón*, uncommon spellings like *.com-Manie*). The remaining 3.32 million tokens can be mapped to 102,215 different lemmas. German has about three times the number of unique lemma types compared to the other languages due to frequent compounds in German.

	Sentences	Tokens	Types
DE	343,620	3,482,804	201,611
EN	166,397	2,077,319	75,272
FR	307,555	3,838,037	113,942
IT	259,868	3,232,256	117,381

Table 2: Corpus size: number of sentences, tokens and word types for all languages in the PDF corpus

	Lemma Types	Unknown Lemmas
DE	102,215	161,644
EN	32,527	70,156
FR	24,494	205,769
IT	24,992	212,824

Table 3: Corpus size: number of unique lemma types (excluding type "unknown") and unknown lemmas (counted per token) for all languages in the PDF corpus

4 Parallel Corpus Alignment

In order to exploit our parallel corpus, we need to compute cross-language alignments on all levels. First, we need to determine document alignments. This is particularly tricky for those Bulletin issues where not all articles were translated. For example, in the 1980s only about half the articles in the parallel German and French issues were translated

```
<corpus>
  <article n="a223" id="cs-2013-03-15-Bali">
        <h1 cat="Society">
          <s n="a223-s1">
                <w n="a223-s1-w1" lemma="a" pos="DT">A</w>
                <w n="a223-s1-w2" lemma="New" pos="NP">New</w>
                <w n="a223-s1-w3" lemma="Life" pos="NP">Life</w>
                <w n="a223-s1-w4" lemma="in" pos="IN">in</w>
                <w n="a223-s1-w5" lemma="Northern" pos="NP">Northern</w>
                <w n="a223-s1-w6" lemma="Bali" pos="NP">Bali</w>
          </s>
        </h1>
        <p class="date">
          <s n="a223-s2">
                <w n="a223-s2-w1" lemma="@card@" pos="CD">05.03.2013</w>
          </s>
        </p>
        <p class="abstract">
          <s n="a223-s3">
                <w n="a223-s3-w1" lemma="for" pos="IN">For</w>
                <w n="a223-s3-w2" lemma="@card@" pos="CD">35</w>
                <w n="a223-s3-w3" lemma="year" pos="NNS">years</w>
        <!-- ... -->
</corpus>
```

Figure 2: XML structure of the corpus

into English and Italian. So, which articles are present across languages? And are they full-length translations or only abbreviated versions?

As a first step towards article alignment, we use the automatically detected article boundaries together with author information. We then check for overlapping names and numbers in the documents. Finally, we compare the article structure (number of paragraphs) and length (in characters) to decide on full-length vs. abbreviated translation. Our first results indicate that in most cases we have full-length translations which makes for a valuable parallel corpus.

Based on the aligned articles we compute sentence alignment. Since the articles in the OCR version and the PDF version contain noise at different places in the text, we need a robust alignment method. We use Bleualign (Sennrich and Volk, 2011) with machine translation of language 1 into language 2 in order to align the sentences in the language 1 text to the corresponding sentences in language 2. The machine translation output is compared with the help of a simplified version of the BLEU metric to the sentences in the language 2 text. This metric, combined with a diagonalization heuristic, results in high precision sentence alignments. Figure 3 shows the final representation of sentence alignments in XML.

Subsequently, word alignment can be performed with GIZA++, Berkeley aligner or any other word aligner of choice. Word alignment will not be included in the corpus release because it is clearly application-specific. Machine translation needs a recall-oriented word alignment while linguistic research requires a high precision word alignment.

5 Corpus Annotation and Lemma Disambiguation

We use the TreeTagger to annotate the corpus with Part-of-Speech (PoS) tags and lemmas in all four languages. The TreeTagger assigns a PoS tag to each token and additionally assigns a lemma if it has seen the token in its training corpus. In case it has seen multiple lemmas for a specific word form, it will assign multiple lemmas.

Since we often find such ambiguous words with multiple lemmas in our German corpus, we investigated two methods for the disambiguation of these lemmas. First, we select among lemma options when we re-attach separated verb prefixes to the lemma. For example, the 1st/3rd person plural verb form *drängen* can have the lemmas *drängen* (EN: to urge) or *dringen* (EN: to insist). The TreeTagger assigns both lemmas to this verb form. If *drängen* occurs with the separated prefix *auf*, then our reattachment algorithm finds that only the combination *aufdrängen* is possible (EN: to force on, to impose), and we remove the other lemma option. This approach is described in more detail in (Volk et al., 2016).

Second, we use the parallel texts for the resolution of these lemma ambiguities. The following example illustrates the advantage of using an English translation to disambiguate German lemmas:

(1) German: *Viele Wege führen nach Rom.*

English: *Many roads lead to Rome.*

The TreeTagger annotates the German word "führen" with the two lemmas "fahren|führen", because when interpreted as a finite verb form, "führen" can be 1st and 3rd person plural in present tense of "führen" (EN: to lead) or a subjunctive form of "fahren" (EN: to go, to drive). However, if the English translation is considered, it becomes clear that the intended meaning in example 1 is "führen". Therefore, we use the parallel articles in English to disambiguate lemmas in the German corpus with the following approach.

First, we computed token alignments between lemmas with GIZA++ over the whole sentence-aligned news part of our corpus (roughly 1.7 million tokens per language) and obtained the lexical translation probabilities in both directions: German to English and English to German. Then, we extracted the German sentences in which a word has multiple lemmas. Using sentence alignment, we retrieved the English translation of each sentence. Next, we searched for the most likely lexical translation of each possible lemmas (as one token) in the English sentence. We checked whether the lexical translation probability from the English lemma to one of the German lemma options is higher than the others. If so, we accepted the more likely lemma and in this way disambiguated the German word.

This is in essence similar to using parallel corpora for word sense disambiguation (as e.g. described in (Shahid and Kazakov, 2013) for Europarl and (Lefever and Hoste, 2014)). Of course, we capture different word senses only if they are reflected in different lemmas. This means that we work more coarse-grained than word sense disambiguation methods which distinguish WordNet senses.

Let us exemplify our algorithm with the above example sentence. We determine the most likely translation of the ambiguous lemma "fahren|führen" by checking the lexical translation probabilities for each token in the parallel English sentence. The pair "fahren|führen – lead" is the top candidate. We then compare the lexical translation probabilities for "fahren – lead" and for "führen – lead". Since the probability for the latter is higher, the lemma "führen" wins.

In this approach, we excluded the combined lemmas "er|es|sie" (which is the lemma that the TreeTagger assigns to the frequent reflexive pronoun "sich") as well as ambiguities due to polite pronoun forms "sie|Sie" (EN: they, you). They cannot be disambiguated reliably with our approach because English has no grammatical gender and no explicit polite forms. We evaluated the quality of our disambiguation system with 100 disambiguated lemmas and reached a precision of 0.97 as shown in table 4 in the first row. However, only about 16% of the lemma ambiguities could be resolved. Therefore, we generalized our approach in order to disambiguate more lemmas.

We were able to improve our recall by assuming that whenever only one of the possible lemmas occurs in the lemma alignments, this should be the disambiguated form. Of course, we cannot avoid creating some false positives with this method, but the number of true positives is far larger. For example, the German word form "Stunden" is ambiguous (EN: "hours" or the nominalized form of "to defer"). The lemma can either be "Stunde" or "Stunden". However, only "Stunde" occurs in our corpus in the lemma alignments on its own. Therefore, all lemmas "Stunde|Stunden" in the corpus are disambiguated to "Stunde". Table 4 shows that with this assumption, we were able to disambiguate about 75% of all ambiguous lemmas. Again, we tested the quality on 100 disambiguated lemmas and reached a precision of 0.93.

As another measure to improve recall, in all cases where no lemma has a higher probability than the others, we checked whether any of the individual lemmas occurs more often in lowercased form. The usefulness of this approach can be observed in the following example:

(2) German sentence:
Es gibt ja den Spruch "Lesen bildet."

English translation:
It is said that: "Reading makes you smarter."

The German word "Lesen" is ambiguous because it could either be the gerund of "to read" but it could also be the plural form of "harvest". The most likely translation of "Lese|Lesen" is the English lemma "reading". Neither the pair "Lese – reading" nor "Lesen – reading" occurs in the lemma alignments in our corpus. Consequently,

```
<linkGrp toDoc="CS_news_corpus_en.xml" fromDoc="CS_news_corpus_de.xml">
        <link type="1-1" xtargets="a96-s1; a1-s1"/>
        <link type="1-1" xtargets="a96-s2; a1-s2"/>
        <link type="1-1" xtargets="a96-s3; a1-s3"/>
        <link type="1-1" xtargets="a96-s4; a1-s4"/>
        <link type="1-1" xtargets="a96-s5; a1-s5"/>
        <link type="1-1" xtargets="a96-s6; a1-s6"/>
        <link type="1-1" xtargets="a96-s7; a1-s7"/>
        <link type="1-1" xtargets="a96-s8; a1-s8"/>
        <link type="1-2" xtargets="a96-s9; a1-s10 a1-s11"/>
        <link type="1-2" xtargets="a96-s10; a1-s12 a1-s13"/>
        <link type="1-1" xtargets="a96-s11; a1-s14"/>
        <!-- ... -->
</linkGrp>
```

Figure 3: XML stand-off annotation for sentence alignments

	Disambig.	Precision
only LTP	16%	97%
LTP & General	75%	93%
LTP & General & Lower	84%	91%

Table 4: Results for automatic lemma disambiguation (LTP = lexical translation probability, General = generalized method, Lower = lemmas also checked in lowercase form). The table shows for each method the percentage of lemmas that we were able to disambiguate and the precision calculated over 100 disambiguated lemmas.

the lemma of the word "Lesen" would remain ambiguous. But when applying the method described above, we also check whether the lowercased German lemmas occur in the word alignments together with "reading". Since the pair "lesen – reading" exists in our corpus and "lese – reading" does not, "Lesen" is accepted as the correct lemma and the word "Lesen" is disambiguated. The results of this refined method can be found in table 4 in the last row. We were able to disambiguate about 84% of the ambiguous lemmas. Tested on 100 disambiguated lemmas, we achieved a precision of 0.91.

We have shown that our lemma disambiguation approach works well for German when using English alignments for the disambiguation. It remains to be investigated how the results change when using French or Italian as parallel texts either alone or in combination. For example, both languages use polite forms which we can exploit to disambiguate even more German lemmas.

Another idea is to use a machine translation system to first translate sentences for which we do not have a parallel text. Then, using our trained lemma alignments from the corpus, we can also disambiguate lemmas for which we do not have translations. We hope to enhance our results further by including more aligned sentences for the training of the lemma alignments. As of now, we evaluated the lemma disambiguation only on the news part of our corpus but we will include the larger Credit Suisse Bulletin corpus in future experiments.

6 Corpus Representation

We store the corpus in a custom XML format[3], see Figure 2 for an example. The outermost element is called "corpus", and it contains a sequence of "article" elements. "article" elements come with an "id" attribute to facilitate document alignment across languages. Within articles, we preserve headers and paragraph structure. Also, we retain information on the general category of the news articles (banking, economy, society, sport etc.). Similar to TEI documents, the "s" element represents a sentence and it contains "w" elements for single words. Each word element has attributes to store lemma and PoS information. We distribute the corpus files together with information on sentence alignments and a document type definition (DTD) with which the validity of the documents can be proven.

We computed the sentence alignment of our corpus with HunAlign in the version that is integrated into InterText (Vondřička, 2014). We store these sentence alignments in separate files as stand-off annotations that link two corpus files, see figure 3.

[3]We will provide an XSLT transformation to convert the corpus into valid TEI documents.

The format allows a quick overview of the alignment types. For example, in the English-German news part of our corpus we have 77,651 sentence alignments of type 1-1 and 8230 alignments of type 1-2, in contrast to 1616 of type 2-1. The aligner also returns zero alignments (2736 0-1 vs. 905 1-0 alignments). There are also a few unbalanced alignments (e.g. 3 times 1-6 alignments and one 7-1 alignment) which indicate omissions in one of the languages.

7 Related Work

This work is a continuation of our efforts to build large diachronic parallel corpora for different text genres. In the past, we have built a multilingual corpus of alpine texts (mountaineering reports, articles about the climate, the geology and geography of the world's mountain regions) (Volk et al., 2010). That corpus is special in terms of genre and also because it spans over 150 years (from 1864 until today), with parallel texts in German and French from 1957 until today (and additional Italian translations since 2012). It also includes untranslated English texts by the British Alpine Club.

Our work is related to others in various ways. Work on building large parallel corpora is, for instance, described in publications by (Steinberger et al., 2006) for the JRC Acquis, by (Lison and Tiedemann, 2016) for OpenSubtitles, and by (Ziemski et al., 2016) for a United Nations corpus over six languages. But there is no literature on building genre-specific diachronic parallel corpora which involve OCR of documents and a longer time span. With respect to the banking domain, our Bulletin corpus is related to the European Central Bank corpus available from the OPUS website[4]. But that corpus is based solely on web site information from recent years.

Regarding applications, there are many papers on using parallel corpora for tasks as diverse as translation studies (Zanettin, 2012) or bilingual terminology extraction, see e.g. (Bertaccini and Tadolini, 2011) and (Macken et al., 2013), not to mention statistical machine translation. Noteworthy are also the various usage-oriented online lexicon systems that are based on parallel corpora and word alignment like Linguee, Glosbe or Multilingwis (Volk et al., 2014, Clematide et al., 2016).

[4]http://opus.lingfil.uu.se/ECB.php

8 Conclusions

We have described the necessary considerations when building a large multilingual corpus based on source documents in various formats (printed, PDF, HTML files). We have shared our experiences with OCR tools and PDF converters. We argued that precise article boundary recognition is central to high quality article alignment across languages. We have suggested a method to semi-automatically detect the article boundaries with high precision.

We also developed two methods for selecting among multiple lemmas which were assigned by the PoS tagger. The first method exploits constraints given by separated verb prefixes. The other, more general method relies on cross-language word alignment and translation probabilities in the parallel corpus.

Our current version of the Bulletin corpus consists of roughly 5 million tokens in each of the three languages French, German, and Italian (1.7 million from the news and 3.3 million from the PDF files) and somewhat less for English. Digitization and corpus building of the issues prior to 1998 is ongoing. We expect to collect a total of 20 million each for French and German by the end of the project.

The corpus is distributed in XML with PoS tags, lemmas and sentence alignments. It is freely available for research purposes.

Acknowledgments

We are grateful to the many students who have contributed to the Bulletin4Corpus project. In particular, we acknowledge valuable contributions by Till Salinger (PDF conversion), Dolores Batinic and Fabienne Leuenberger (sentence alignment), Dominique Sandoz (web crawling), and Katrin Affolter (processing of the Abbyy OCR output).

We would like to thank Credit Suisse for their consent that the Bulletin texts can be made available for language technology research. We also acknowledge valuable support by the library of the Swiss National Bank in Zurich and the Swiss National Library in Berne.

References

Franco Bertaccini and Marianna Tadolini. 2011. Banking terminology: creation of a terminology database Italian-German. In *Proceedings of the First International Conference on Terminology, Languages, and Content Resources*, Seoul.

Simon Clematide, Johannes Graën, and Martin Volk. 2016. Multilingwis – a multilingual search tool for multi-word units in multiparallel corpora. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives / Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*. Tradulex, Geneva.

Hervé Dejean. 2015. Extracting structured data from unstructured documents with incomplete resources. In *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 271–275, Nancy.

Els Lefever and Véronique Hoste. 2014. Parallel corpora make sense: Bypassing the knowledge acquisition bottleneck for word sense disambiguation. *International Journal of Corpus Linguistics*, 19(3):333 – 367.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Lieve Macken, Els Lefever, and Veronique Hoste. 2013. Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology*, 19(1):1–30.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of The 18th International Nordic Conference of Computational Linguistics (Nodalida)*, Riga.

Ahmad R. Shahid and Dimitar Kazakov. 2013. Using parallel corpora for word sense disambiguation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 336–341, Hissar, Bulgaria.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Carmelia Ignat, Tomaz Erjavec, Dan Tufiş, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC*, Genoa.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *Proceedings of LREC*, Valletta, Malta.

Martin Volk, Lenz Furrer, and Rico Sennrich. 2011. Strategies for reducing and correcting OCR errors. In C. Sporleder, A. van den Bosch, and K. Zervanou, editors, *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing, pages 3–22. Springer-Verlag, Berlin.

Martin Volk, Johannes Graën, and Elena Callegaro. 2014. Innovations in parallel corpus search tools. In *Proceedings of LREC*, Reykjavik.

Martin Volk, Simon Clematide, Johannes Graën, and Phillip Ströbel. 2016. Bi-particle adverbs, PoS-tagging and the recognition of German separable prefix verbs. In *Proceedings of KONVENS*, Bochum.

Pavel Vondřička. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Federico Zanettin. 2012. *Translation-driven corpora: corpus resources for descriptive and applied translation studies.* St. Jerome Publishing.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.